Министерство образования Республики Беларусь

Учреждение образования

«Белорусский государственный университет информатики и радиоэлектроники»

Кафедра интеллектуальных информационных технологий

**Лабораторная работа № 2**

«Методыавтоматического распознавания языка текстового документа»

Вариант 26

|  |  |
| --- | --- |
| Выполнили студенты  группы 721702: | Гурбович А.И.  Каленик П.Р. |
| Проверил: | Крапивин Ю.Б. |

МИНСК

2020

**Цель работы** – изучить и отработать практические навыки применения методов автоматического распознавания языка текстовых документов.

**Вариант:**

|  |  |  |  |
| --- | --- | --- | --- |
| 26 | Испанский, Немецкий | Html | Частотных слов, алфавитный, ваш метод |

## Информация о тестовой коллекции документов:

* Тестовые документысформированы на испанском и немецком языках в формате .html

### Пример тестового документа на немецком языке:

### Имя файла: german\_sample.html

### Тип файла: HTML-документ

### Размер: 4КБ

### Содержание:

Juliana kommt aus Paris. Das ist die Hauptstadt von Frankreich. In diesem Sommer macht sie einen Sprachkurs in Freiburg. Das ist eine Universitätsstadt im Süden von Deutschland. Es gefällt ihr hier sehr gut. Morgens um neun beginnt der Unterricht, um vierzehn Uhr ist er zu Ende. In ihrer Klasse sind außer Juliana noch 14 weitere Schüler, acht Mädchen und sechs Jungen. Sie kommen alle aus Frankreich, aber nicht aus Paris. Julianas beste Freundin Marie macht auch gerade einen Sprachkurs, aber in Hamburg, das liegt ganz im Norden von Deutschland. Wenn die beiden ihre Schule beendet haben, wollen sie in Deutschland studieren. Juliana will Tierärztin werden, ihre beste Freundin auch. Aber Maries Eltern sind beide Zahnärzte, deshalb wird Marie wahrscheinlich auch Zahnärztin werden. Juliana und Marie verbringen insgesamt sechs Wochen in Deutschland. Nach dem Sprachkurs machen sie eine Prüfung.

### Пример тестового документа на испанском языке:

* + Имя файла: spanish\_sample.html
  + Тип документа: HTML-документ
  + Размер: 4КБ
  + Содержание:

Mark está de viaje de negocios en Barcelona. Hoy tuvo un día libre y salió a visitar la ciudad. Primero, caminó por La Rambla, la calle más famosa de Barcelona, llena de gente, tiendas y restaurantes. Se dirigió al Barrio Gótico, uno de los sitios más antiguos y bellos de la ciudad. En la Plaza Sant Jaume observó dos de los edificios más importantes: El Palacio de la Generalitat de Catalunya y el Ayuntamiento. Volvió a La Rambla. Mark tenía hambre y se detuvo a comer unas tapas y beber una cerveza. Continuó hasta la grande y hermosa Plaza de Catalunya. Avanzó por el Paseo de Gràcia hasta llegar a un edificios fuera de lo común Casa Batlló y luego a Casa Milà, diseños del arquitecto Antoni Gaudí. Quiso saber más sobre este famoso arquitecto y se dirigió al Park Güell, donde tomó muchas fotografías. El día se acababa pero antes de volver al hotel, Mark tomó un taxi hacia la Fuente Mágica y disfrutó de un espectáculo de agua y luces. Mark quedó sorprendido con esta gran ciudad y sintió que le faltó tiempo para conocer más lugares interesantes. Se prometió regresar para tomar unas vacaciones con su familia.

## Описание структур данных, использованных для хранения входной и выходной информации:

### Входная информация:

Текст на web-странице, на немецком либо испанском языке.

### Выходная информация:

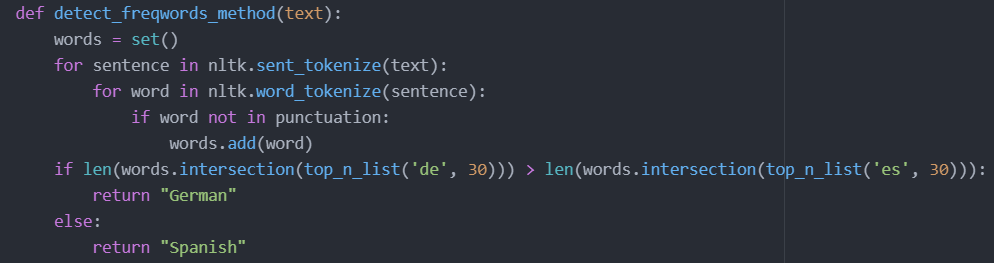
Информация выводится в консоль для наглядного отображения пользователю, так же при желании, пользователь может сохранить результат идентификации текста каждого отдельного метода в текстовый документ.

Текстовый документ для результата идентификации текста - result.txt

## Описание алгоритмов определения языка(в текстовом и графическом виде):

### Метод частотных слов:

Программа берёт 30 самых распространённых немецких и испанских слов используя библиотеку Wordfreq и проверяет из какой группы (немецких или испанских слов) больше в тексте.



### Алфавитный метод:

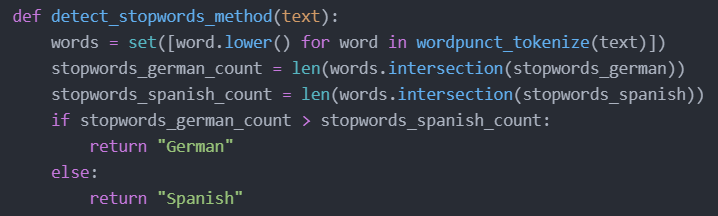
У немецкого и испанского языка есть спец символы.

Программа разбивает текст из HTML-страницы побуквенно и заносит их всех во множество. По итогу язык текста выбирается на основе пересечения с наибольшей мощностью между алфавитом немецких символов и испанских символов.



Свой метод:

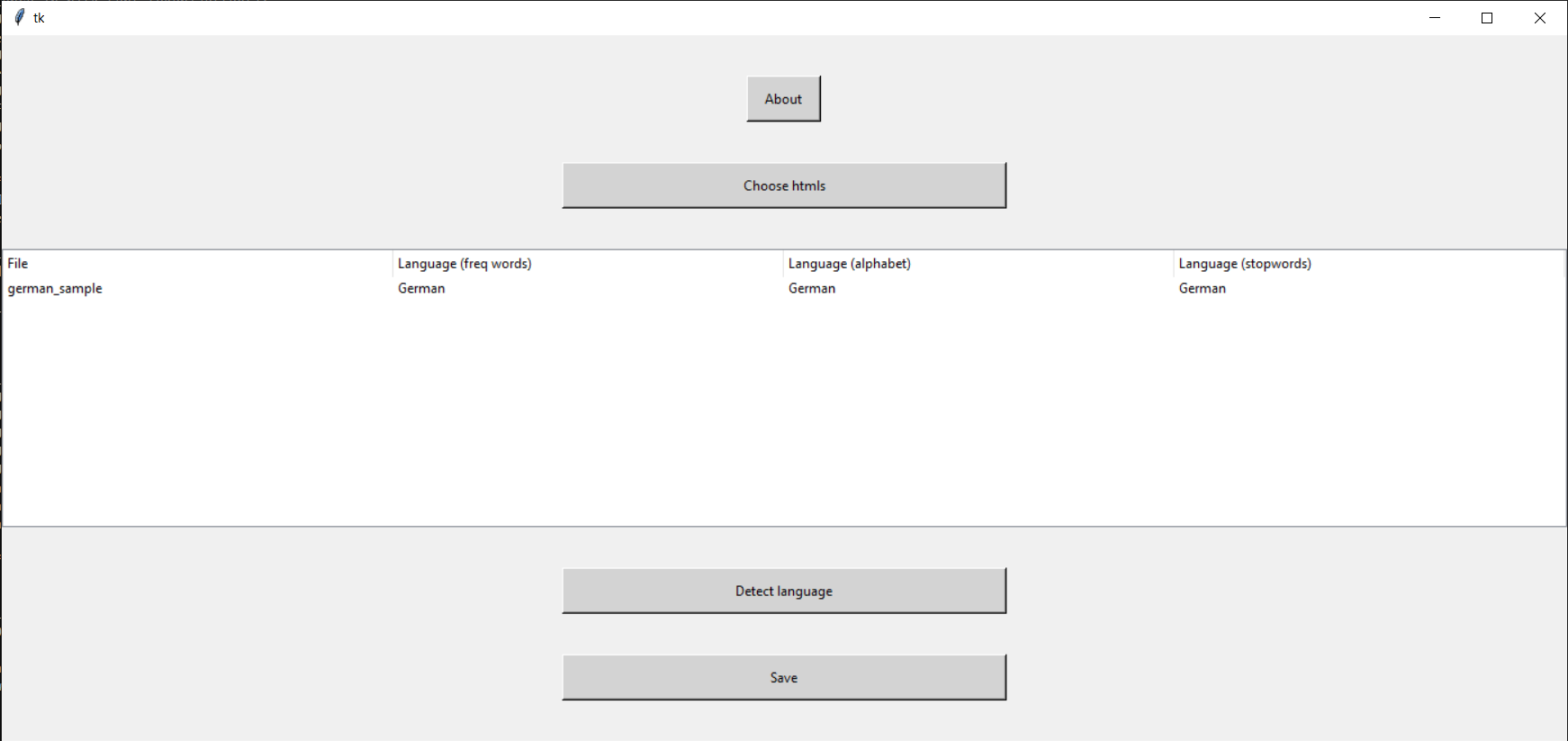
Программа использует библиотеку nltk со стоп словами. Язык определяется на основе того, стоп слов какого типа (немецкого или испанского языка) больше в тексте.

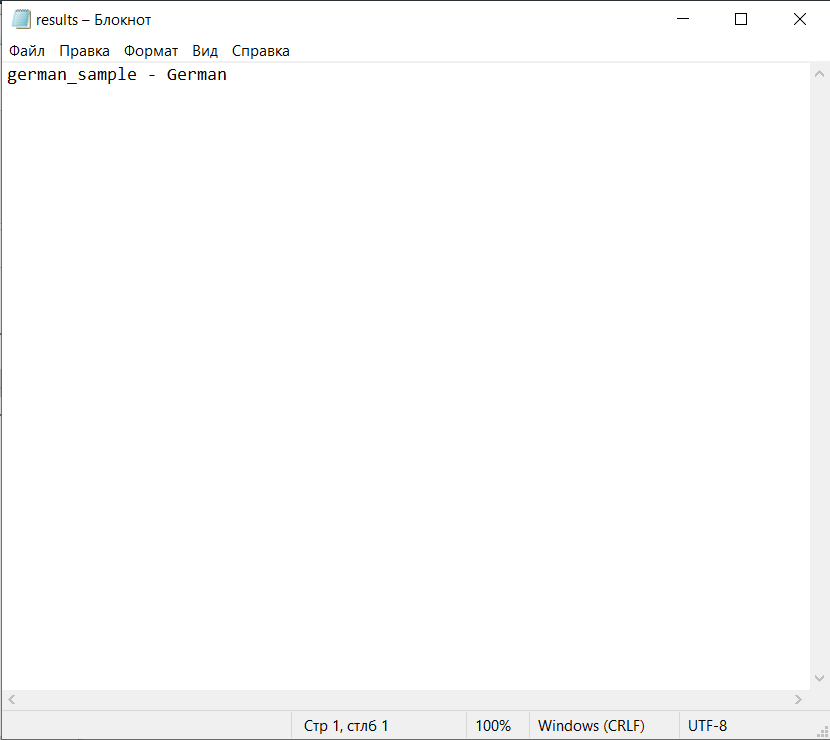


## Результаты тестирования системы:

Стартовый экран  


Входная информация: german\_sample.html

Проверка проводится сразу по тремя методами.  
Результат работы программы:  


Файл result.txt после нажатия на кнопку Save имеет следующее содержимое:  


## Описание и особенности применения готовых к использованию компонент:

В данной системе использовались сторонние библиотеки:

Wordfreq - библиотека на языке python, использована для наиболее распространённых слов в испанском и немецких языках..

nltk - библиотека на языке python, использована для своего метода со стоп словами.

tkinter- библиотека на языке python, использована для реализации пользовательского интерфейса.

## Вывод:

В результате выполнения лабораторной работы, разработана система по определению принадлежности текста к немецкому или испанскому языку, с возможностью сохранения выходной информации. Методы разработанные в системе, в результате тестирования системы на разных текстах, показали хороший результат и высокую точность в определении языка текста. В настоящее время система работает только с немецким и испанским языками, но предусмотрены все возможности по добавлению новых языков в систему и обучению системы на этих языках.